Improving our statistical inferences requires meta-research

Robert T. Thibault[1,2], Marcus R. Munafò[1,2]

[1]School of Psychological Science, University of Bristol
[2]MRC Integrative Epidemiology Unit at the University of Bristol

Please address correspondence to robert.thibault@bristol.ac.uk (https://orcid.org/0000-0002-6561-3962)

word count: 912
keywords: meta-research; statistical significance; clinical significance; p-value; statistical inference; reproducibility; replication

In the past few years, there has been renewed interest in the use of thresholds for declaring statistical significance and, if we decide to use thresholds, what alpha value is optimal. Should we lower the threshold for claiming statistical significance from $p < .05$ to $p < .005$[1]? Should we justify the selected alpha level for every study separately[2]? Or should we abandon statistical significance altogether[3]? Whilst these proposals may appear at odds, they align in terms of their goal—to improve the robustness of our inferences. However, empirical data that sheds light on the potential costs and benefits of these approaches remains scarce.

Koletsi and colleagues[4] attempt to address this issue. They focus on the first proposal—reducing the threshold for declaring statistical significance from $p < .05$ to $p < .005$—and provide empirical evidence for the impact of doing so. They show that lowering the statistical significance threshold in meta-analyses of clinical interventions would not substantially reduce the number of true discoveries. In their sample, only 2.9% (95% CI 0.4-10%) of *recommended* interventions would be disregarded (i.e., incorrectly excluded) if the more stringent $p < .005$ threshold was applied, whilst 39% (95% CI 24-57%) of interventions that have some statistical evidence (i.e., $p < 0.05$) but are currently *recommended against* or *not recommended* for other reasons would be disregarded (i.e., correctly excluded). In other words, the benefits of lowering the threshold for claiming statistical significance in meta-analyses of clinical interventions outweigh the costs.

This study forms part of a growing literature investigating the research process itself (sometimes described as meta-research, or research on research). In an ideal world, researchers would iteratively identify and investigate problems in the research process, develop potential solutions, and evaluate these[5]. This approach has the potential to provide a background process of continual improvement in how we work. Elements of this approach are certainly not new. Researchers have identified shortcomings in the use of statistical inference in published research for at least a half century, investigated them more thoroughly over the past decade or two, and recently suggested a range of potential solutions (e.g., [1–3]). What remains scarce, however, is the *evaluation* of potential solutions. Koletsi and colleagues address this gap.

The evaluation step of meta-research provides the empirical evidence necessary to justify changes we might wish to make to the research ecosystem. For example, meta-research studies have shown how journal data sharing policies may increase data *availability*, but leave the *usability* of the shared data

low[6], that reporting checklists may not actually improve compliance with guidelines[7], and that preregistration of study protocols might allow questionable research practices to be more readily identified, but does not always reduce these practices[8]. Each of these studies evaluates a specific implementation of a proposed solution or improvement. Without this evidence, we would be left to simply implement what *seem* like good ideas (but may not be in practice). And that is clearly inadequate.

Koletsi and colleagues focus on a specific proposed statistical reform in the context of meta-analyses and clinical practice recommendations: reducing the threshold for claiming statistical significance. In some cases, such as when deciding whether or not to recommend a clinical intervention, dichotomous inference may be necessary. In this context, the results presented by Koletsi and colleagues provide support for a pragmatic revision of the conventional alpha level, from 0.05 to 0.005. The data show, however, that relying solely on a revised alpha level is no panacea: 15% (95% CI 7-25%) of *recommended* interventions are lost to the revised threshold and 24% (95% CI 16-33%) of the interventions that meet the revised threshold are either *recommended against* or *not recommended*.

Crucially, Koletsi and colleagues only arrive at their minimal cost of a 2.9% reduction in *recommended* interventions by going beyond a strict statistical threshold and considering additional factors. This (perhaps ironically) illustrates that, for many research questions, the dichotomisation of statistical evidence is simply too crude an approach, at least on its own. Koletsi and colleagues are in effect combining proposed solutions: they are using alpha values to dichotomise the evidence and then drawing on additional information for borderline decisions.

We can also improve our statistical inferences through other initiatives largely independent of how we define and use statistical thresholds. Preregistering hypotheses and analysis plans can safeguard against selective reporting and p-hacking—or at least provide the evidence to demonstrate when these practices have not been done. The Registered Reports submission format offered by many journals allows for peer-review of methods before conducting the research[9]. This places peer review earlier in the research process than is usual, and at a stage where it can impact the design of a study, not just the reporting. Open research practices such as sharing data, analysis code, and research materials can also provide transparency, greater checking, and perhaps improved quality control.

The conversation on improving research practices has thus far focused largely on the epidemiology of the problem—what are the factors that contribute to poor quality research. We are now seeing research into *how* to make the transition to better research practice—what are the interventions, how can we implement them, and what is the evidence that they work. This effort requires coordination among stakeholders across the research landscape: funders, publishers, institutions, policy makers, and researchers[10]. An investment in research that evaluates solutions and studies how best to implement them will allow for iterative improvements in how we do research, and will return that investment many times over.

Code

We calculated the four confidence intervals reported in this commentary using the R code *binom.test(x, n, alternative="two.sided", conf.level=0.95)* where *x* and *n* come from the *Primary analysis* section of Koletsi and colleauges[4] and are 2/68, 15/38, 10/68, and 23/97.

Koletsi and colleagues coded intervention recommendations based on information available on UpToDate—a point-of-care medical resources that provides recommendations for clinical practice[11].

Acknowledgements
We thank members of the ReproducibiliTea Journal Club at the University of Bristol for insightful discussions related to this commentary. Robert Thibault is support by a postdoctoral fellowship from the *Fonds de la recherche en santé du Québec*.

Conflicts of interest
Both authors report no conflicts of interest.

References

1. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nat Hum Behav* [Internet]. Springer US; 2018;**2**(1):6–10. Available from: http://dx.doi.org/10.1038/s41562-017-0189-z
2. Lakens D, Adolfi FG, Albers CJ, et al. Justify your alpha. *Nat Hum Behav* [Internet]. Springer US; 2018;**2**(3):168–171. Available from: http://dx.doi.org/10.1038/s41562-018-0311-x
3. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon Statistical Significance. *Am Stat*. 2019;**73**(sup1):235–245.
4. Koletsi D, Solmi M, Pandis N, Padhraig S. Fleming CUC, Ioannidis JPA. Most recommended medical interventions reach P<0.005 for their primary outcomes in meta-analyses. *Int J Epedimiology*. 2019;
5. Hardwicke TE, Serghiou S, Janiaud P, Danchev V. Calibrating the scientific ecosystem through meta-research. *Annu Rev Stat its Appl*. 2019;
6. Hardwicke TE, Mathur MB, Macdonald K, et al. Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Preprint*. 2018;1–72.
7. Hair K, Macleod MR, Sena ES. A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARus). *Res Integr Peer Rev*. Research Integrity and Peer Review; 2019;**4**(1):1–17.
8. Goldacre B, Drysdale H, Dale A, et al. COMPare: A prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*. Trials; 2019;**20**(1):1–16.
9. Chambers CD. Registered Reports: A new publishing initiative at Cortex. *Cortex*. 2013;
10. Munafò MR, Nosek BA, Bishop DVM, et al. A manifesto for reproducible science. *Nat Hum Behav* [Internet]. Macmillan Publishers Limited; 2017;**1**(1):1–9. Available from: http://dx.doi.org/10.1038/s41562-016-0021
11. UpToDate. Grading Guide [Internet]. 2019 [cited 2019 Nov 18]. Available from: https://www.uptodate.com/home/grading-guide